XVI. Magyar Számítógépes Nyelvészeti Konferencia Szeged, 2020. január 23–24.

Automatikus ékezetvisszaállítás transzformer modellen alapuló neurális gépi fordítással

Laki László János[1,2], Yang Zijian Győző[1,2]

[1]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
[2]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
1083 Budapest, Práter u. 50/a.
{laki.laszlo, yang.zijian.gyozo}@itk.ppke.hu

Kivonat Cikkünkben egy ékezetvisszaállító programot mutatunk be, amelyet a jelenkori „state-of-the-art" transzformer modellen alapuló neurális gépi fordító rendszerrel tanítottunk be. A mobil eszközökön történő üzenetírás elterjedésével és a minél gyorsabb szövegbevitelre való törekvéssel tömeges jelenséggé vált az ékezetes betűk elhagyása a gépelt írásban. Ennek egyik következménye, hogy a interneten elérhető – főleg a szociális médiából származó – korpuszok egy része ékezetmentes. Egy ékezetvisszaállító program segítségével vissza tudjuk állítani az ékezethiányos szavakat, valamint integrálva szövegbeviteli eszközökkel támogatni tudjuk a felhasználók számára a szövegbevitelt. Az általunk létrehozott rendszer, annak ellenére, hogy semmilyen morfológiai elemzőt nem használ, több mint 99,7%-os pontossággal tudja helyesen visszaállítani az ékezeteket magyar nyelv esetében. A hibaanalízis során kiderült, hogy a hibák több mint 50%-a a többértelműségből fakad, illetve, hogy a rendszerünk által ajánlott ékezetesítés utáni mondat is helyes. Készítettünk egy demó felületet is, amelyen ki lehet próbálni a különböző modellek működését.
Kulcsszavak: ékezetvisszaállítás, neruálisháló-alapú gépi fordítás, NMT, transzformer modell

1. Bevezetés

Napjaink számítógépes nyelvészei számára nagy lehetőséget nyújtanak az interneten elérhető nagy mennyiségű szövegek. Számos részterületen használjuk a weboldalakról összegyűjtött korpuszokat, mint például a gépi fordítás, a szöveg kivonatolás vagy az érzelem detektálás. Ezekhez a feladatokhoz viszont nélkülözhetetlen, hogy a vizsgált szöveg a lehető legjobb minőségű legyen.

A mobil eszközökön írt szövegek és üzenetek esetében tömegjelenséggé vált az ékezetes betűk elhagyása. Ennek következményeképp léteznek olyan korpuszok is, amelyek egy része ékezetmentes, így nem működnek rajtuk a természetes szövegen betanított szövegfeldolgozó modellek. Egy ékezetvisszaállító program segítségével vissza tudjuk állítani az ékezethiányos szavakat, valamint integrálva további szövegbeviteli eszközökkel támogatni tudjuk a felhasználók szövegbevitelét. A feladat komplexitását a többértelmű szavak visszaállítása is növeli.

Az elmúlt években a neurálishálózat-alapú módszerek eredményei túlszárnyalták az addigi legjobb rendszereket. Ez a nyelvtechnológia területén is megmutatkozik, ezért célunk az volt, hogy megvizsgáljuk az ékezetesítés problémáját a jelenlegi „state-of-the-art" NMT-alapú rendszerrel.

2. Kapcsolódó munkák

Az elmúlt években több kísérlet is született az ékezetek helyreállítását megcélozva. Nyelvfüggetlen módszerekkel kísérletezett Mihalcea és Nastase (Mihalcea és Nastase, 2002). Kutatásukban gépi tanulásos módszereket alkalmaztak a probléma megoldására. Az egyik módszer, amikor az ékezetes betűk pozíciója és környezete segíti a megoldást. Ezzel a megközelítéssel 95%-os pontosságot értek el. Egy másik módszerükkel korpuszból becsülték meg a különböző ékezetes szavak disztribúcióját, mellyel 98%-os pontosságot értek el. A rendszer hátránya viszont, hogy a korpuszban nem szereplő – ismeretlen szavakat – nem tudja kezelni.

Charlifter szintén nyelvfüggetlen megoldást keresett (Scannell, 2011). Lexikonalapú statisztikai módszerekkel állítja helyre az ékezetet. Figyeli a közvetlen környezetet és az ismeretlen szavak kezelésére karakteralapú statisztikai modellt alkalmaz. A legjobb esetben is csak 93%-os pontosságot ért el.

Nyelvspecifikus kutatásokat végzett Yarowsky spanyol és francia nyelvre (Yarowsky, 1999), valamint Zweigenbaum és Grabar francia nyelvre (Zweigenbaum és Grabar, 2002).

Magyar nyelvre Németh és társai (Németh és mtsai, 2000) egy text-to-speech alkalmazást mutatnak be, melyben kezelik az ékezethiányos szavakat. A probléma megoldásához morfológiai és szintaktikai elemzőt is használnak, mellyel 95%-os pontosságot értek el. Novák és Siklósi (Novák és Siklósi, 2015, 2016) statisztikai gépi fordítást (SMT) alkalmaznak az ékezet helyreállításához. Morfológiai elemző nélküli és egy morfológiai elemzővel rendelkező SMT-vel is végeztek kísérleteket. A legjobb eredményt – 99,06% – a morfológiai elemzővel érték el.

Nagy Péter szakdolgozatában (Nagy, 2018) RNN-alapú neurális gépi fordítórendszert alkalmazott a feladat megoldására. Legjobb eredménye eléri a 99,5%-os pontosságot. Munkája során BPE (Byte pair encoding) tokenizálást (Sennrich és mtsai, 2015) végzett úgy, hogy külön modellt használt a forrás- és a célnyelvi tanítóanyagra. Ez a dolgozat tekinthető a leghasonlóbbnak saját munkánkhoz. Kutatásunkban az RNN modell helyett a jelenlegi „state-of-the-art" transzformer modellt használjuk, míg a BPE helyett a közös szótárral rendelkező Sentence Piece (SPM) tokenizálást alkalmazunk. Ezzel a technikával sikerül további javulást elérnünk.

3. Ékezetes szavak helyreállítása

A korpusz-alapú gépi fordító rendszer lényege, hogy transzformációt képez tetszőleges forrás- és célnyelvi mondatok között, ahol a rendszer betanításához nem

kell más, mint egy kétnyelvű párhuzamos korpusz. Az ékezetes szavak helyreállítására kézenfekvő választás a gépi fordítás módszereit használni, mivel az ékezetlen és az ékezetes mondatok grammatikailag, szókincsileg és szószerkezetileg nagyon hasonlóak.

A neurális hálózat tanításához nagy mennyiségű tanítóanyagra van szükség, melynek előállítása a jelen feladathoz igen könnyű. A tanítóanyag létrehozásához annyit kellett tenni, hogy egy egynyelvű korpusz ékezetes karaktereiről szabály alapon eltávolítottuk az ékezeteket.

3.1. A korpusz

A korpusz létrehozásához az online elérhető Open Subtitles[1] nevű angol-magyar párhuzamos korpuszának magyar oldali szövegét használtuk. A korpusz TV és mozi filmekre létrehozott feliratokból áll. Ennek megfelelően főleg rövidebb, informális mondatokat tartalmaz. A gépi fordító rendszer célnyelvi korpuszának előállításához a mondatokban az ékezetes karaktereket lecseréltük az ékezet nélküli párjukra (pl. á→a, é→e stb.)

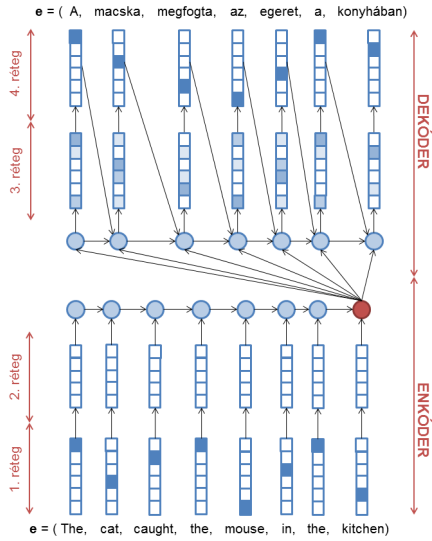
A korpusz megközelítőleg 29 millió szegmensből áll, melyből 5000 mondatot validációs és 3000 mondatot tesztelési célra elkülönítettünk. A korpusz az egyik legnagyobb szabadon hozzáférhető párhuzamos tanítóanyagnak számít, ellenben mérete elmarad az egynyelvű tanítóanyagokétól. Választásunk azért esett erre az adathalmazra, mert több párhuzamos kutatásunk során is használjuk, és néhány koprusztisztító lépést már előzetesen eszközöltünk rajta. Kivettük azokat a mondatokat, amelyek speciális karaktereket (pl. kínai, japán, cirill stb.) tartalmaztak, valamint a teszt halmaz mondatait kézzel kijavítottuk. Mérete elégséges a neurális hálózatok helyes betanítására, valamint a tanítási idő is viszonylag kezelhető marad (1-2 nap).

3.2. A neurális gépi fordítórendszer

A 2010-es évek első felére a statisztikai gépi fordítórendszerek elérték teljesítőképességük határát. Az alapjait képező módszert és a létrehozott keretrendszereket a kutatók nagyon sok befektetett munka ellenére lényegében nem sikerült tovább javítani. Az áttörést (Bahdanau és mtsai, 2015) rendszere hozta el, ami egy figyelmi modellel támogatott enkóder-dekóder architektúrájú NMT rendszer volt. A modell lényege, hogy kettéválasztja a fordítás folyamatát két elkülöníthető részre. A kódolás során lényegében egy RNN-alapú seq2seq modellt hoz létre, tehát a szóbeágyazási modellhez hasonlóan a fordítandó modellekből egy n-dimenziós vektort készít. Az 1. ábrán ez a vektor felel meg az ábra közepén látható piros/sötét node-nak. A második fázis a dekódolás, ahol a mondatvektorból generálja ki a célnyelvi mondatot egy RNN réteg segítségével.

Innentől számítva az NMT rendszerek átvették a vezető szerepet az SMT-től. 2017-ben a Google cég munkatársai (Vaswani és mtsai, 2017) publikálták és szabadon hozzáférhetővé tették az úgynevezett multi-attention réteggel támogatott

[1] http://opus.nlpl.eu/OpenSubtitles-v2018.php



1. ábra: Enkóder-dekóder architektúra vázlatos rajza

NMT rendszerüket. Ezt a szakirodalomban transzformer-alapú architektúrának nevezik. A módszer lényege, hogy az eddigi egy helyett több figyelmi réteget helyeztek el a rendszerben, ami segítségével nagymértékben nőtt a többértelmű szavak fordításának minősége.

Munkánk során a Marian NMT(Junczys-Dowmunt és mtsai, 2018) nevű keretrendszert használtuk, ami egy c++ nyelven íródott szabadon hozzáférhető programcsomag. Könnyen telepíthető, jól dokumentált, memória- és erőforrás-optimális implementációjának köszönhetően [2] az akadémiai felhasználók és fejlesztők által leggyakrabban használt eszköz (Barrault és mtsai, 2019).

3.3. A Sentence Piece tokenizáló

Az NMT rendszerek működése GPU processzorokon történik, melyek egyik szűk keresztmetszete a bennük található memória mérete. Ez határozza meg a létrehozható NMT rendszer szótárának a méretét. Egy szóalapú rendszer esetében az általánosságban 100K különálló szóban korlátozzák le a rendszert, így a további szavakat ismeretlenként kezeli.

(Sennrich és mtsai, 2015) ezt a problémát úgy oldották meg, hogy a szavak helyett úgynevezett subword (szótöredék) szintre csökkentették a legkisebb fordítási egységet. A BPE (Byte Pair Encoding) egy adattömörítő eljárás, ahol a leggyakoribb bájtpárokat egy olyan bájttal helyettesítjük, amely nem szerepel

[2] https://marian-nmt.github.io/

magában az adatban. Az eljárás a korpuszon először egy karakteralapú szótárat hoz létre, ahol minden szót karakterek sorozataként ábrázol. Ezután gyakoriság alapján a gyakori karaktersorozatokat önálló tokenekként kezeli. Ezzel az adat tömörítése mellett az ismeretlen szavak kezelését is megoldja, hiszen a részszavakból előállítható egy olyan összetétel, amely nem szerepelt eredetileg a korpuszban.

Ezt a módszert fejlesztették tovább (Kudo és Richardson, 2018). Az általuk létrehozott Sentence Piece nevű eszköz egy felügyelet nélküli szöveg tokenizáló és detokenizáló, melyet elsősorban a neurálishálózat-alapú gépitanulásos feladatokhoz fejlesztettek ki. Implementálva van benne a BPE metrika, ami egy unigram nyelvmodellel (Kudo, 2018) van súlyozva. Használatával elhagyhatók a költséges nyelvspecifikus előfeldolgozási lépések, mint például a tokenizálás vagy a kisbetűsítés. A módszer lényege, hogy a természetes szöveget úgy alakítja át, hogy abban a különböző „szavak" száma korlátos legyen, valamint az így létrejött tanítóanyagban nem lesznek ismeretlen szavak. Ennek köszönhetően a neurális hálózatok paraméterszáma nagymértékben csökkenthető.

(1) Sima szöveg: Petőfi Sándor egy nagyszerű költő.
SPM szöveg: Pető fi □ S ándor □egy □nagyszerű □költő .

A fenti példában látható az SPM modell kimenete. A sima szöveg szavait gyakran előforduló karakter sorozatokra tördeli szét. Érdekes megfigyelni, hogy az eredeti mondat szóközeit is a szavakhoz csatolja és mint önálló karaktert (□) kezeli.

3.4. Megjelenítő felület

Készítettünk egy demó felületet[3], amelyen ki lehet próbálni a különböző modellek működését. Egy lenyíló menün keresztül lehet kiválasztani a tesztelendő modellt, majd egy input mezőn begépelhetjük a szavakat. A szóközök után megvizsgálja az addig leírt szövegrészletet, és ha hibásnak találja, ajánlatot tesz a javításra.

4. Kísérletek

A minőségbeli összehasonlíthatóság végett betanítottuk (Novák és Siklósi, 2015) által leírt morfológiai elemző nélküli SMT-t, valamint (Nagy, 2018) által jegyzett RNN-alapú neurális gépifordító rendszert (NMT-RNN) is.

Az SMT tanításához a Moses nevű keretrendszert (Koehn és mtsai, 2007) használtuk, ahol a nyelvmodellt a KenLM-el (Heafield, 2011) hoztuk létre. A rendszer tanítása során az alapbeállításokat használtuk és kihagytuk a szóösszerendelő és az átrendező lépéseket. Ezekre a lépésekre nem volt szükségünk, hiszen azonos a szószám és a szósorrend monoton a forrás- és a célnyelvi oldalon.

[3] http://nlpg.itk.ppke.hu/projects/accent

A fordítás előfeldolgozó fázisában történik egy tokenizáció és egy „truecase" lépés. A truecase-ing egyfajta kisbetűsítés, ahol a mondat kezdő szaváról döntjük el, hogy azt alapesetben kis- vagy nagybetűs formában használjuk. A fordítás utófeldolgozása során történik egy „detruecase" lépés és egy detokenizáció.

Az NMT tanításához a Marian neurális gépi fordítórendszert használtuk. Az RNN-alapú NMT beállításhoz a (Nagy, 2018) dolgozatában szereplő értékeket vettük alapul. A rendszer fontos jellemzője, hogy a BPE modellt külön tanítja be a forrás- és a célnyelvi korpuszokból. Ennek az a jelentősége, hogy a két esetben ugyanazt a szót a forrás- és a célnyelvi oldalon különböző subword formában tördelheti a rendszer. Emiatt sérül a szavak egy-egyértelmű megfeleltethetősége. A rendszer másik tulajdonsága, hogy 5 egymást követő sikertelen iteráció után megáll (early-stopping).

Munkánk elméleti alapját az újonnan elérhető transformer és SPM technológiák adták. Kíváncsiak voltunk arra, hogy ékezetesítés esetén is sikerül-e elérni a rendszerek természetes nyelveken bemutatott minőségi javulását. A továbbiakban NMT-TM néven hivatkozunk a saját modellünkre. A rendszerünk tanításához az alábbi paramétereket használtuk:

- Sentence Piece: szótárméret: 16000; egy szótár a forrás- és egy a célnyelvi korpusznak; karakter lefedettség a teszt korpuszon: 100%
- Transformer modell: enkóder és dekóder rétegeinek száma: 6; transformer-dropout: 0,1;
- learning rate: 0,0003; lr-warmup: 16000; lr-decay-inv-sqrt: 16000;
- optimizer-params: 0,9 0,98 1e-09; beam-size: 6; normalize: 0,6
- label-smoothing: 0,1; exponential-smoothing

5. Eredmények

Kutatásunk során megmértük a gép által adott szóalapú eredmény pontosságát (precision), fedését (recall) és az abszolút pontosságot (accuracy). Mivel a gépi fordítás során az eredetileg helyes szavak is megváltozhatnak, szükséges megvizsgálni a fordítás pontosságát az összes szóra (ALL). Ezenkívül elvégeztük a kiértékelést azokra a szavakra nézve is, amelyek rendelkeznek magánhangzóval (MGH), vagyis a feladatra nézve releváns szavakat.

	ALL			MGH		
	Pontosság (Precision)	Fedés (Recall)	Abs pontosság (Accuracy)	Pontosság (Precision)	Fedés (Recall)	Abs pontosság (Accuracy)
SMT	97,96%	96,97%	98,49%	98,13%	97,04%	98,49%
NMT-RNN	97,04%	97,54%	98,58%	97,16%	97,60%	98,56%
NMT-TM	**99,38%**	**99,28%**	**99,63%**	**99,42%**	**99,33%**	**99,62%**

1. táblázat. A különböző ékezetesítő modellek kiértékelése

A 1. táblázat eredményei alapján láthatjuk, hogy az általunk létrehozott rendszer, amely transzformer modellt és Sentence Piece tokenizálót használ, min-

den esetben a legjobb eredményt ért el. Érdemes megemlíteni, hogy annak ellenére, hogy pontosságban (precision) az SMT jobb eredményt ért el az NMT-RNN modellhez képest, fedésben (recall) és a rendszer pontosságát (accuracy) illetően az NMT-RNN teljesített jobban.

Az NMT-TM modell teljesítménye minden esetben meghaladja a 99,27%-ot. Összesített rendszerszintű pontossága eléri a 99,63%-ot.

6. Hibaelemzés

Az eredmények mélyebb elemzése során megvizsgáltuk a rendszer által elkövetett hibákat. A 2. táblázatban láthatjuk a hibatípusokat. A tesztanyag 3000 mondatából (18438 token és 8957 type) mindössze 67 mondatban volt hiba, melyekben összesen 69 darab szót vélt hibásnak. Ezen hibák mélyebb elemzéséből láthatjuk, hogy nagy részük nem tekinthető valódi hibának. Az egyik ilyen hibakör a szövegkörnyezet ismerete nélküli többértelműségből származó szavak esete.

(2) REF: Különben nem hoznák haza.
RES: Különben nem hoznak haza.

A másik típus az azonos jelentésű, de különböző alakú szavak elrontása. Ezeket az eseteket szintén nem számoljuk hibásnak. Ha ezeket az eseteket nem tekintjük hibáknak, akkor a rendszerünk **99,83%-os** relatív pontosságot ér el.

(3) REF: Hova mész nyaralni?
RES: Hová mész nyaralni?

A hibáknak kevesebb mint fele valódi hiba, de a valódi hibák fele a tulajdonnevek helyesírásából fakad. Ezt azért fontos megemlíteni, mert ha a gép soha sem látott példát egy-egy tulajdonnév helyesírására, akkor nem is várhatjuk el tőle, hogy tökéletesen ékezetesítse azt.

Hibatípus	**Arány (db)**	**Példák** (referencia (ref) - eredmény (res))
Helyes kimenet	**55,07% (38 db)**	
Ekvivalens alakok	7,26%	hova - hová, tied - tiéd
Értelmes kimenet	92,74%	ref: Érdekelné ez a dolog? res: Érdekelne ez a dolog? ref: Különben nem hoznák haza. res: Különben nem hoznak haza.
Valódi hibák	**44,93% (31 db)**	
Tulajdonnevek	45,16%	Liúról - Liuról, Ramával - Rámával
Hibás értelmezés	54,84%	még - meg, melyen - mélyen, teli - téli

2. táblázat. A Tranformer modell különböző hibatípai

A Transformer modell teljesítményének egyik érdekes eredménye, hogy kizárólag csak ékezettel kapcsolatos hibákat vétett, de ez az RNN modell vagy a MOSES esetén nem így volt. A 3. táblázatban látható az RNN modell és a MOSES azon hibatípusai, amelyek eltérnek a Transformer modelltől.

Modell	Hibatípus	Példa (referencia - eredmény)
RNN, MOSES	Kis- és nagybetű	Átvehetitek - átvehetitek; Azt - azt
	Más formátum	7:54 - 7: 54; a"... - a..."; Szóval, ... - Szóval,..... Gary's - Gary 's
MOSES	Nem ékezesített	széttörnéd - szettorned; hőtermelés - hotermeles; túlárasztotta - tularasztotta

3. táblázat. Az RNN modell és MOSES hibatípusai, amelyek eltérnek a Tranformer modelltől

7. Összegzés

A kutatásunkkal létrehoztunk egy ékezetvisszaállító rendszert. A rendszer tanításához egy neruálishálózat-alapú gépi fordítórendszert használtunk, amely transzformer modellt és Sentence Piece tokenizálót használ. A rendszerünk 99,63%-os pontossággal tudja helyesen visszaállítani az ékezeteket. Végeztünk hibaelemzést is és azt állapítottuk meg, hogy a hibák közel fele nem is valódi hiba, ha ezeket az eseteket helyes kimeneteknek tekintjük, akkor a rendszerünk 99,83%-os pontosságot ér el. A rendszerünkhöz készítettünk egy demó felületet is, amelyen ki lehet próbálni a különböző modellek működését.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással az FK 125217 számú projekt keretében az FK 17 pályázati program, valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1NKP-2018-00008 azonosítójú projekt keretében valósult meg.

Hivatkozások

Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (szerk.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1409.0473

Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., MĂźller, M., Pal, S., Post, M., Zampieri, M.: Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 1–61. Association for Computational Linguistics, Florence, Italy (August 2019), http://www.aclweb.org/anthology/W19-5301

Heafield, K.: KenLM: faster and smaller language model queries. In: Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation. pp. 187–197. Edinburgh, Scotland, United Kingdom (July 2011), https://kheafield.com/papers/avenue/kenlm.pdf

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. pp. 177–180. ACL '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007)

Kudo, T.: Subword regularization: Improving neural network translation models with multiple subword candidates. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 66–75. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018), https://www.aclweb.org/anthology/D18-2012

Mihalcea, R., Nastase, V.: Letter level learning for language independent diacritics restoration. In: Proceedings of the 6th Conference on Natural Language Learning - Volume 20. pp. 1–7. COLING-02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), https://doi.org/10.3115/1118853.1118874

Nagy, P.: Magyar nyelvű zajos szövegek automatikus normalizálása. Szakdolgozat, Pázmány Péter Katolikus Egyetem (2018)

Németh, G., Zainkó, C., Fekete, L., Olaszy, G., Endrédi, G., Olaszi, P., Kiss, G., Kis, P.: The design, implementation, and operation of a hungarian e-mail reader. International Journal of Speech Technology 3(3), 217–236 (Dec 2000), https://doi.org/10.1023/A:1026567216832

Novák, A., Siklósi, B.: Automatic diacritics restoration for Hungarian. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language

Processing. pp. 2286–2291. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015), https://www.aclweb.org/anthology/D15-1275

Novák, A., Siklósi, B.: Ékezetek automatikus helyreállítása magyar nyelvű szövegekben. XII. Magyar Számítógépes Nyelvészeti Konferencia pp. 49–58 (2016)

Scannell, K.P.: Statistical unicodification of african languages. Language Resources and Evaluation 45(3), 375 (Jun 2011)

Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. CoRR abs/1508.07909 (2015), http://arxiv.org/abs/1508.07909

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (szerk.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Yarowsky, D.: A Comparison of Corpus-Based Techniques for Restoring Accents in Spanish and French Text, pp. 99–120. Springer Netherlands, Dordrecht (1999)

Zweigenbaum, P., Grabar, N.: Accenting unknown words in a specialized language. In: Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain. pp. 21–28. Association for Computational Linguistics, Phildadelphia, Pennsylvania, USA (Jul 2002)